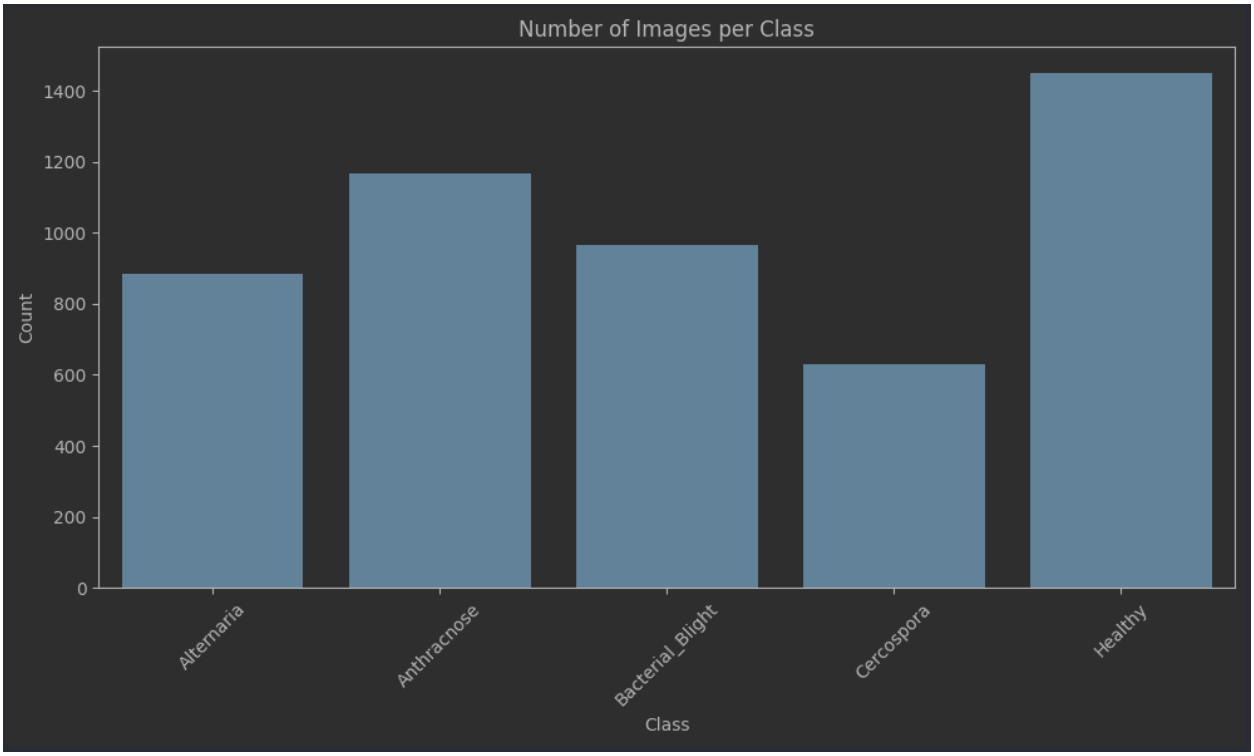
Assignment L02

1. Seturi de date folosite

Am folosit setul de date pentru rodii https://www.kaggle.com/datasets/sujaykapadnis/pomegranate-fruit-diseases-dataset. Acesta are suficiente imagini si are 5 clase difreite: Healthy, Alternaria, Anthracnose, Bacterial_Blight, Cercospora.



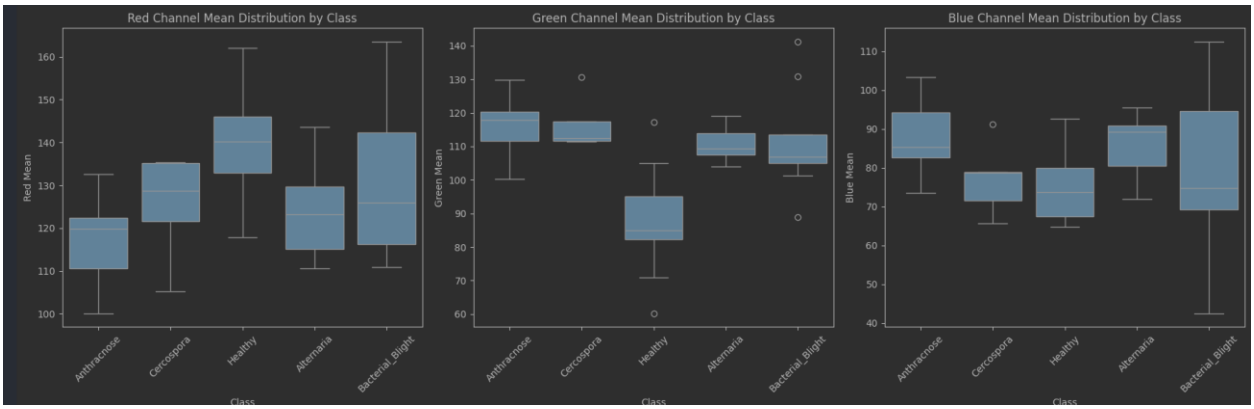
1.Description of data:
Poze in format RGB, dimensiune 3120x3120.

```
Image Properties Summary:
        width  height  channels  file_size_kb
count   100.0   100.0     100.0    100.000000
mean   3120.0  3120.0       3.0    866.661709
std       0.0     0.0       0.0    124.783217
min    3120.0  3120.0       3.0    632.756836
25%    3120.0  3120.0       3.0    778.292236
50%    3120.0  3120.0       3.0    852.405273
75%    3120.0  3120.0       3.0    956.967529
max    3120.0  3120.0       3.0   1170.654297
```

RGB summary:

```
Pixel Statistics Summary:
           r_mean      g_mean      b_mean      r_std      g_std      b_std  \
count   50.000000   50.000000   50.000000  50.000000  50.000000  50.000000
mean   128.817492  105.345129   80.654784  63.790114  46.499708  51.198590
std     14.700172   16.417352   13.375329  10.424078   6.721824   7.458597
min    100.116284   60.220193   42.496253  36.660708  33.057793  36.982398
25%    116.644428   95.142364   72.161947  57.149703  41.142862  45.233270
50%    129.960940  108.439360   80.059494  63.216589  46.318399  49.826172
75%    139.640895  116.338369   89.678339  69.979634  51.574827  57.589286
max    163.520603  141.226826  112.449119  89.619053  58.146618  67.267210
```

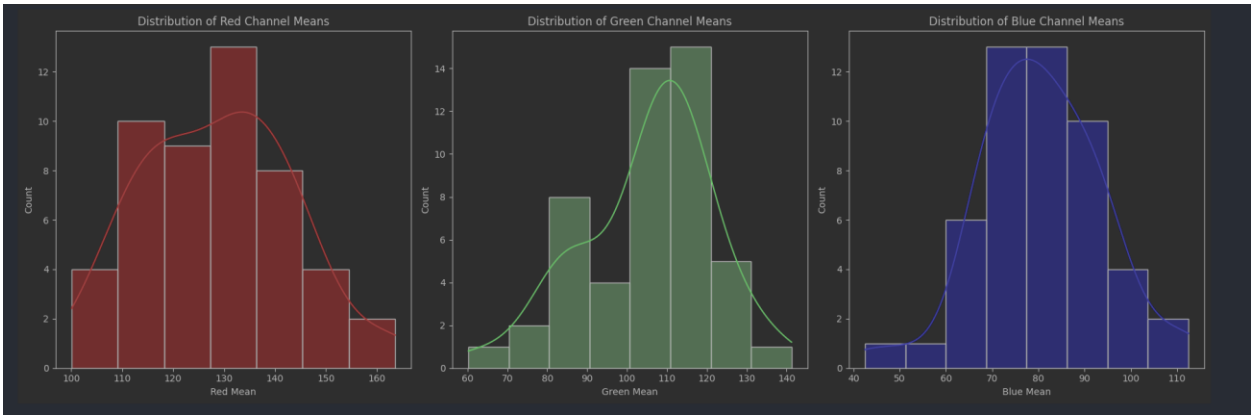
```
Mean Pixel Statistics by Class:
                    r_mean      g_mean     b_mean      r_std      g_std  \
label
Alternaria      123.532158  110.994894  86.302408  59.890938  53.787587
Anthracnose     117.127383  115.906278  87.265437  63.859242  45.768919
Bacterial_Blight  130.063901  111.274627  78.832142  54.429060  50.175065
Cercospora      125.231271  116.755342  77.191809  56.940697  46.306589
Healthy         140.247032   88.005311  75.040588  73.341912  40.895655
```



2.Handling missing data.

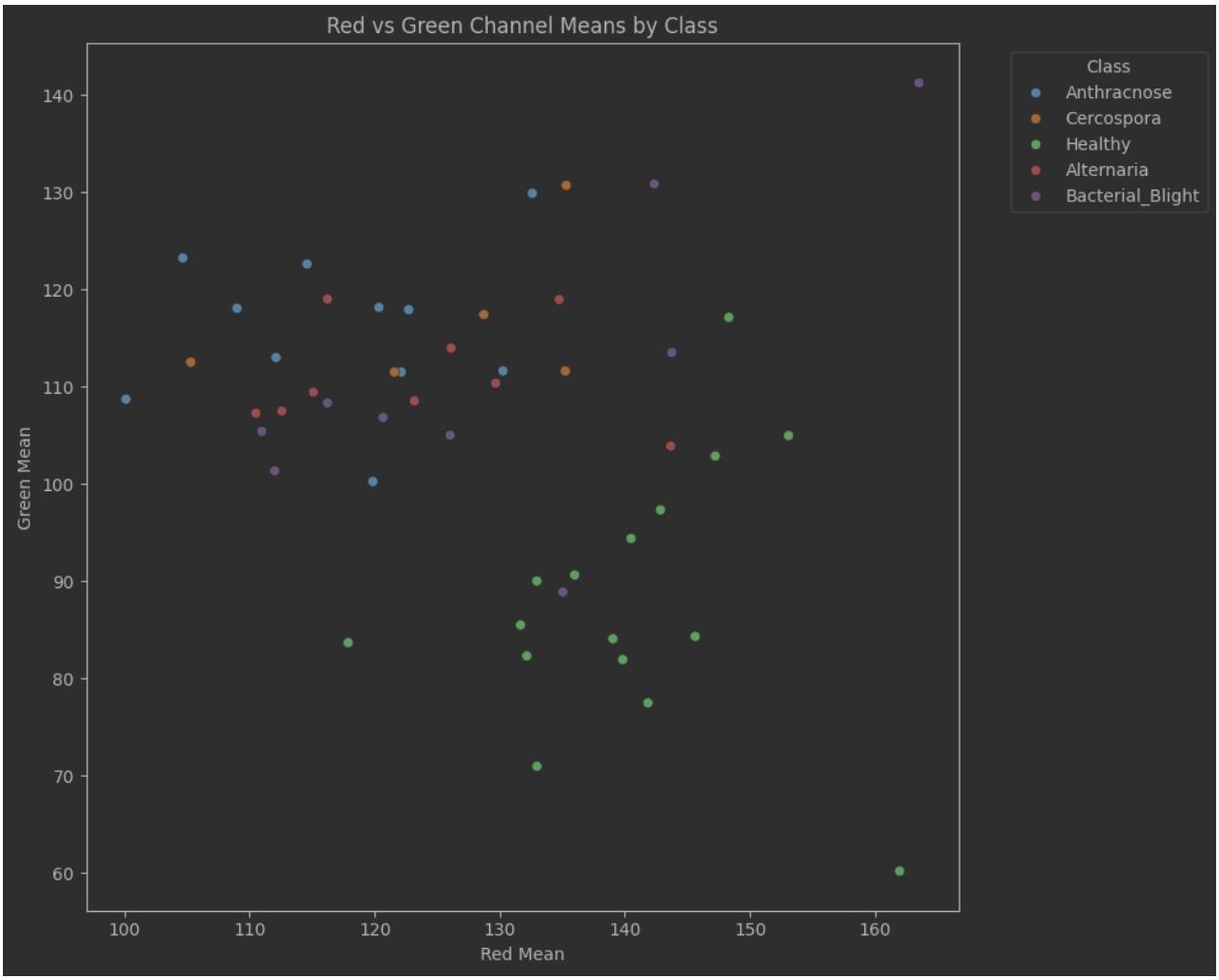
Am folosit atat librariile PIL cat si opencv pentru a verifica integritatea imaginilor. In acest dataset nu am gasit probleme cu nicio imagine. In cazul in care am fi gasit, le-am fi dat drop din Dataset

3.Outliers

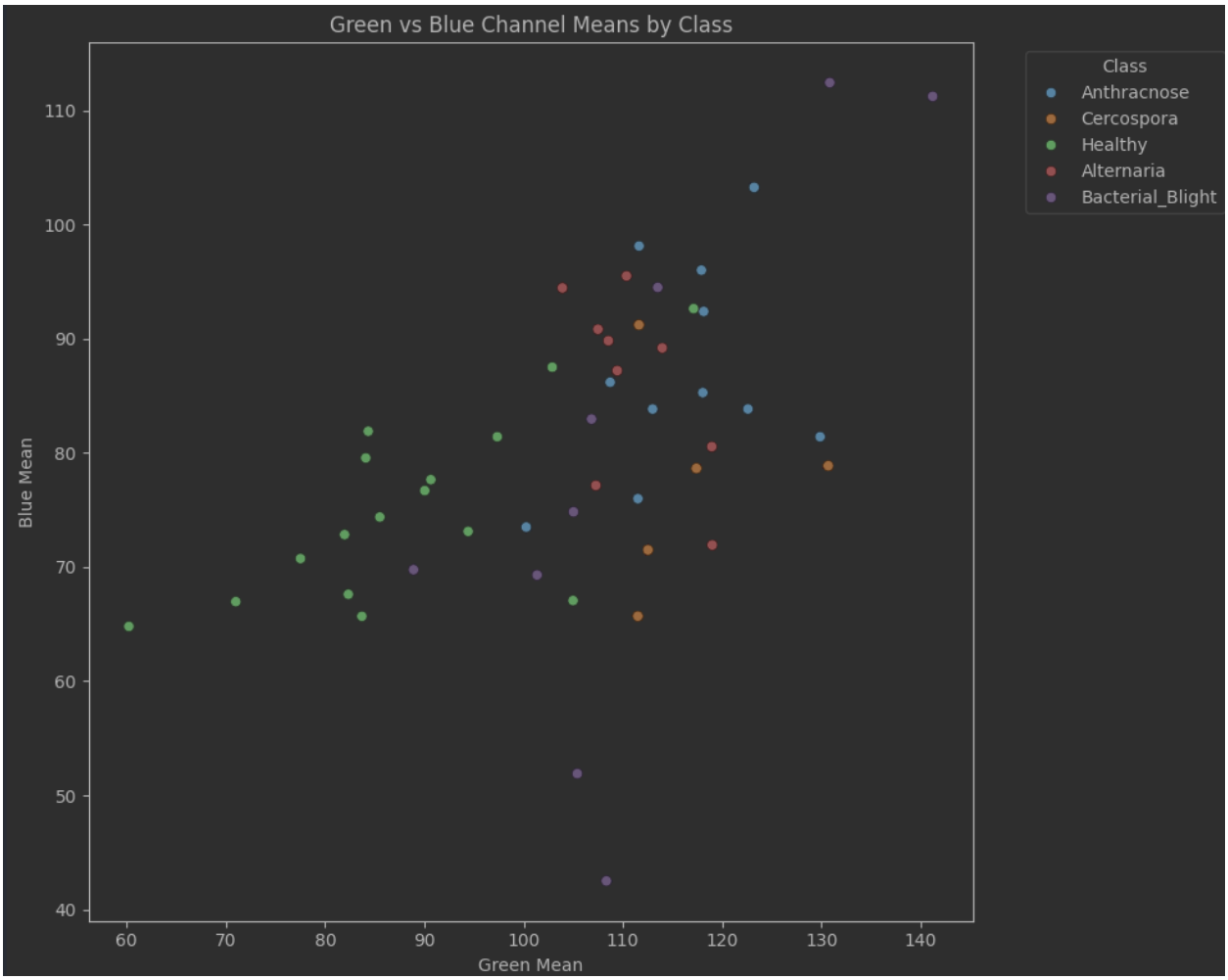


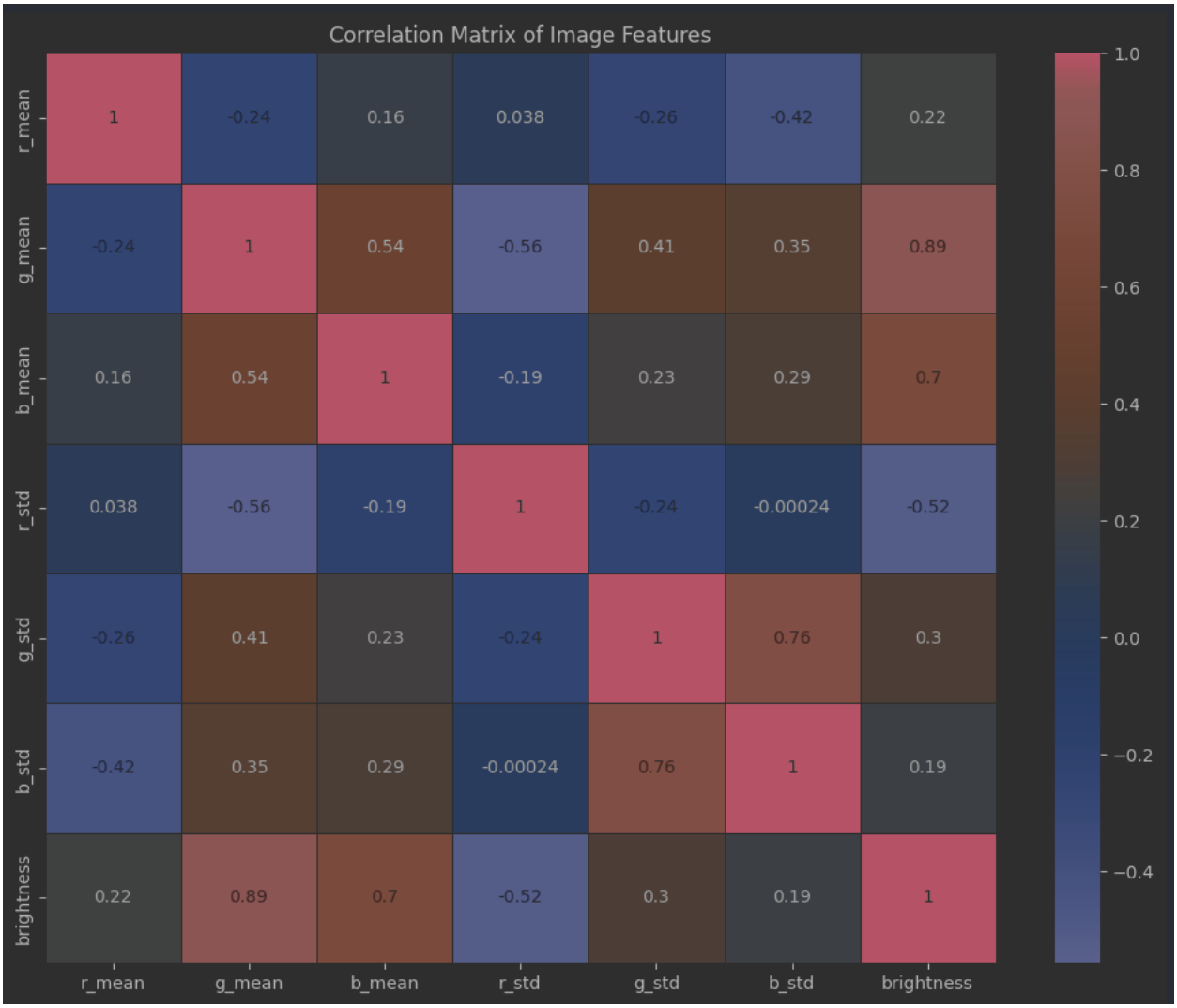
Folosind metoda Z-score am gasit un outlier care nu avea brightness corespunzator. Folosind metoda IQR, am gasit 1 outlier green, 1 outlier blue, si 1 outlier brightness. Toate au fost scoase din dataset.

4.Understanding relationships and new insight through plots.



Se observa o separare a clasei Healthy, avand valori mai mari pentru rosu, si mai mici pentru verde. Clasele cu boli au tendinte mai mari de verde


Green vs Blue Channel Means by Class
Class
Anthracnose
Cercospora
Healthy
Alternaria
Bacterial_Blight
Blue Mean
Green Mean
110
100
90
80
70
60
50
40
60
70
80
90
100
110
120
130
140



Parametrii de brightness si valorile medii ale canalelor verde si albastru sunt puternic corelate

Preprocesarea setului de date.

1. Standardizare

Am testat StandardScaler, MinMaxScaler si RobustScaler. Avand in vedere modifcarile date de Standard si Robust, am decis sa mergem mai departe cu MinMax, pentru pastrarea distributiei relative a pixelilor. In intervalul 0-1.

2.Normalizare

Avand in vedere folosirea MinMax si posiblia pierdere a datelor de culoare, normalizarea datelor devine redundanta.

3.Econding

Am folosint encoding pentru a codifica fiecare clasa intr-un numar.

Dupa analiza datelor, am ales sa mergem cu imaginile in format RGB, redimensionate si standardizate folosind minmax 0-1